Association Mining Based Approach to analyze COVID-19 response and Case Growth in the United States

Satya Katragadda[1], Raju Gottumukkala[1,*], Ravi Teja Bhupatiraju[1], Azmyin Md. Kamal[1], Vijay Raghavan[1], Henry Chu[1], Ramesh Kolluru[1], and Ziad Ashkar[1]

[1]Informatics Research Institute, University of Louisiana at Lafayette, Lafayette, 70506, USA
[*]raju.gottumukkala@louisiana.edu

ABSTRACT

Containing the COVID-19 pandemic while balancing the economy has proven to be quite a challenge for the world. We still have limited understanding of what combination of policies have been most effective in flattening the curve and preventing a rise in infection rate due to the dynamic and evolving nature of the pandemic, lack of quality data, and limited understanding of the relationship between the combination of policies and the infection growth. The paper introduces a novel data mining-based approach to understand the effects of different non-pharmaceutical interventions in containing the COVID-19 infection rate. Association rule mining approach was used to perform descriptive data mining on publicly available data for 50 states in the United States to understand the similarity among various policies and underlying conditions that led to the transition between different infection growth curve phases. Multi-peak logistic growth model is used to label the different phases of infection growth phases. The common trends in the data were analyzed with respect to lockdowns, face mask mandates, mobility, and infection growth. Our observations from this analysis indicate that face mask coverings combined with mobility reduction through moderate stay-at-home orders were most effective in reducing the number of COVID-19 cases across various states.

Introduction

COVID-19 is a viral disease with no effective cure. Until the COVID-19 vaccine became available recently, several non-pharmaceutical interventions were used to contain the outbreak, which included stay-at-home orders, social distancing at 6 feet, limited gatherings, hand washing, refraining from touching the face, and masking. Among these, stay-at-home orders potentially carried a high economic cost in lost revenue and financial support to the unemployed[1]. Given the complex dynamics of COVID-19, the high variability of intervention strategies, and the complexity of pandemic behavior, understanding the differential impact of combinations of various measures is non-trivial.

Various researchers used agent-based simulations and statistical correlation analysis to understand how the spread of COVID -19 was affected by a combination of different policies.[2]. For example, Li et al.[3] used a compartmental model to evaluate the effect of social distancing and cloth face coverings on the spread of infections. In another study, Tatapudi et al.[4] presented a study on Miami-Dade County to understand how social-mixing behavior, SAH orders, and contact tracing affect both the case growth and economy. Other similar efforts include agent-based simulation developed by Silvia et. al.[5], and Ghaffarzadegan[6]. Correlation and regression-based analysis were performed by Badr et al.[7] and Sarmadi et al.[8] to understand the impact of various factors such as mobility on infection spread. This paper introduces an association mining approach to analyze similarities across various policies and infection rates in communities.

Association rule mining (ARM) is a common data mining technique used to discover similarity and dissimilarity among objects[9]. The approach was originally designed to obtain insights into consumer buying habits, such as understanding the groups of products customers would buy together[9]. The approach later garnered interest in many domains[10–13]. In public health, for instance, in recent work, ARM was used to analyze the relationship between environmental stressors and adverse human health impacts[14].

We adapted the ARM approach to gather insights into how various non-pharmaceutical interventions contributed to infection growth. Rather than offer clear hypothesis-based objectives, the proposed technique provides insights into similarities and dissimilarities among the combination of policies and local conditions that led to an increase or decrease in infection rates. We use publicly available data collected from all 50 states to discover common patterns with respect to similarities between six different factors, namely stay-at-home-orders, face masks, population density, mobility, infection rates, on the future infection rates across various states in the United States.

Data and Methods

Association mining allows us to perform a descriptive analysis of patterns between various factors known to influence infection growth rate and the actual infection growth rate. We specifically look at population density, infection rate, face mask orders, stay-at-home orders, and mobility[7,15–18].

Association Rule Mining

Given a dataset containing a collection of records or transactions, each record comprises a set of categorical attributes. One of the attributes is the target attribute of interest. The association rule may be denoted by $A \Rightarrow B$, where A (is the antecedent or LHS) and B (the Consequent or RHS) are sets of various attribute-value pairs (also called itemsets), where A and B are disjoint. The rule represents the hypothesis that when variables in A occurs in the dataset, the variables in B also occurs. Association mining generates a large number of rules from a given dataset. In a dataset with m attributes ($n-1$ antecedents and one consequent) with n values, each can generate a maximum of $nm^(n-1)-1$ rules. However, not all rules are significant. The goal of this approach is to find rules that have high practical significance. To eliminate spurious rules, we define four measures: support, confidence, and lift.

Given two disjoint sets of attribute value pairs A and B, and a association rule $A \Rightarrow B$; support of the rule refers to the number of records where the attribute value pairs in either set A or B appear in the dataset relative to the total number of records (transactions or instances). This denotes the prevalence of the rule in the dataset. By definition the support value is symmetric (i.e. support of both rules $A \Rightarrow B$ and $B \Rightarrow A$ are equal). The confidence of the rule $A \Rightarrow B$, measures the conditional probability of B, given A. Confidence measure for a given rule is asymmetric.

$$support(A \Rightarrow B) = support(A \cup B) = \frac{|set\ of\ records\ containing\ A \cap\ set\ of\ records\ containing\ B|}{total\ number\ of\ transactions} \tag{1}$$

$$confidence(A \Rightarrow B) = \frac{support(A \cup B)}{support(A)} \tag{2}$$

$Lift$ is the ratio between the observed support and the expected support between the independent variables A and B. A $Lift > 1$ implies a greater degree of dependence whereas a $Lift < 1$ indicates negative dependence, and $Lift = 1$ shows that A and B are independent. Lift is also a asymmetric measure between the itemsets A and B.

$$lift(A \Rightarrow B) = \frac{support(A \cup B)}{support(A) \times support(B)} \tag{3}$$

In addition to Lift, chi-squared test has also been used to measure the significance of a rule[19,20]. The chi-squared value of a rule $A \Rightarrow B$ is defined as:

$$\chi^2(A \Rightarrow B) = n(lift(A \Rightarrow B) - 1)^2 \frac{support(A \Rightarrow B)confidence(A \Rightarrow B)}{(confidence(A \Rightarrow B) - support(A \Rightarrow B))(lift(A \Rightarrow B) - confidence(A \Rightarrow B))} \tag{4}$$

where n is the total number of transactions in the dataset. A rule is considered significant if the the chi-squared value is greater than a threshold determined by the number of factors in the antecedent[21].

In this paper, we model the contributing factors (i.e., the antecedent) to be face-covering orders, social distancing orders, mobility, population density, case level, and current incident phase. The target variable (the consequent) is the future incident growth phase. One of the critical assumptions/requirements for ARM is that all the attribute values are discrete. The numerical data used in the study (i.e., mobility, number of cases per capita) were discretized into five quantiles. We discretize the infection growth curve, which is continuous data into 5 phases based on the logistic growth model.

Table 1. Breakdown of all the attributes, their values, and the frequency of the attribute value pairs

Attribute	Data	Attribute Values	Frequency of Attribute Value Pairs
Mask Mandate	No Mask	No Mask	367
Mask Mandate	Countywide	Countywide	179
Mask Mandate	Statewide	Statewide	505

Mask Mandate	Recommended	Recommended	9
Social Distancing	Phase 0	Phase 0	229
Social Distancing	Phase 1	Phase 1	152
Social Distancing	Phase 2	Phase 2	156
Social Distancing	Phase 3	Phase 3	277
Social Distancing	Phase 4	Phase 4	226
Social Distancing	Phase 5	Phase 5	20
Mobility Levels	0 - 15%	Very Low	212
Mobility Levels	>15% - 40%	Low	217
Mobility Levels	>40% - 50%	Medium	209
Mobility Levels	>50% - 60%	High	216
Mobility Levels	>60%	Very High	206
Population Density	0 - 150	Low	609
Population Density	>150 - 300	Medium	273
Population Density	>300	High	168
Cases per Capita	0 - 0.1%	Very Low	220
Cases per Capita	>0.1% - 0.3%	Low	220
Cases per Capita	>0.3% - 1%	Medium	239
Cases per Capita	>1% - 2%	High	203
Cases per Capita	>2%	Very High	178
Current Incidence Phase	Early Growth	Early Growth	286
Current Incidence Phase	Fast Growth	Fast Growth	438
Current Incidence Phase	Decline	Decline	222
Current Incidence Phase	Steady state	Steady state	95
Current Incidence Phase	End state	End state	9
Future Incidence Phase	Early Growth	Early Growth	313
Future Incidence Phase	Fast Growth	Fast Growth	419
Future Incidence Phase	Decline	Decline	231
Future Incidence Phase	Steady state	Steady state	80
Future Incidence Phase	End state	End state	7

Data collection & Preprocessing

Our study includes weekly aggregated data from all the 50 states within the United States between June 1st and November 15th, 2020. We start our data collection on June 1st because including earlier data may skew our analysis (only 8 states had a mask mandate before June and most of the states were under lockdown[22]). We end our study period on November 15th before the start of the winter holiday season. Discretized attributes, values, and the frequency distribution of each attribute-value pair are presented in Table 1.

Mask Usage

We used the official face-covering orders issued by various governors or local authorities from AARP State-by-State Guide to Face Mask Requirements[23] and Masks4All compilation[24]. We rounded the dates to the start of the workweek. The four categories of mask orders are No-Mask, Countywide, recommended (statewide), and mandated (statewide). The discretized dataset we produced and detailed definitions of each of these orders were provided on GitHub[25]. The state mask mandate variation across all the states is illustrated in Figure 1.

State Reopening

All states initiated a strict lockdown at the beginning of the pandemic in March 2020. The states modified these orders based on the perceived risk of cases, hospitalizations, and deaths while also trying to bring back the economy. States mostly adapted the guidelines provided by the White house COVID-19 task force reopening procedures[26,27]. The specific orders that were considered include Phase-0, Phase-1, Phase-2, Phase-3, Phase-4, and Phase-5. Detailed definitions of each of these orders were provided at his webpage[25].

Mobility Levels

The mobility information was from the Descartes Labs, a popular dataset used by several studies for analyzing the relationship between mobility and COVID-19 case growth[5,28,29]. The dataset uses anonymized mobile device locations to calculate a

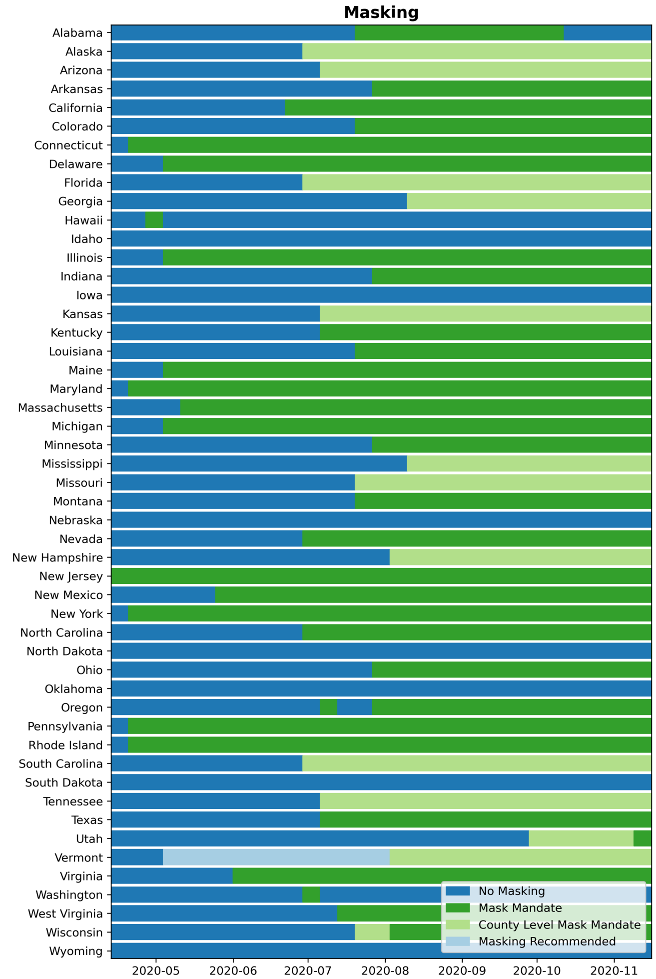

Figure 1. Various mask mandates issued across all the states across the United States

local mobility metric. The metric represents the median of the max-distance traveled by individuals at state and county level normalized to the metric before the pandemic[30].

Population Density

The population density of each state represents the number of people per square mile of land area based on the estimated 2020 population[31].

Cases per capita

The official COVID-19 weekly case data from June 1st to November 10th for the United States was extracted from JHU[32] . The per capita cases were calculated based on the estimated 2019 US Census population data.

Incidence Phases

The incidence growth rate of the pandemic is discretized into five phases based on the standard intervals obtained from a logistic growth curve[33,34]. Given the states have multiple peaks, we use a multi-peak-based logistic growth model from Batista et al.[33] to obtain discrete phases. Phase-I is called the *early growth phase* (or ascending) where (b) Phase-II is the *fast growth phase* which falls between the end of the lag phase (or slow growth phase) and the peak (c) Phase-III is the *decline phase* where the cases decrease from fast growth to steady-state, (d) Phase-IV – *steady-state* and finally (e) Phase-V is the *ending phase*. The first 4 phases are illustrated for the state of Arizona in Figure 2, the fifth phase is not visible in the image.

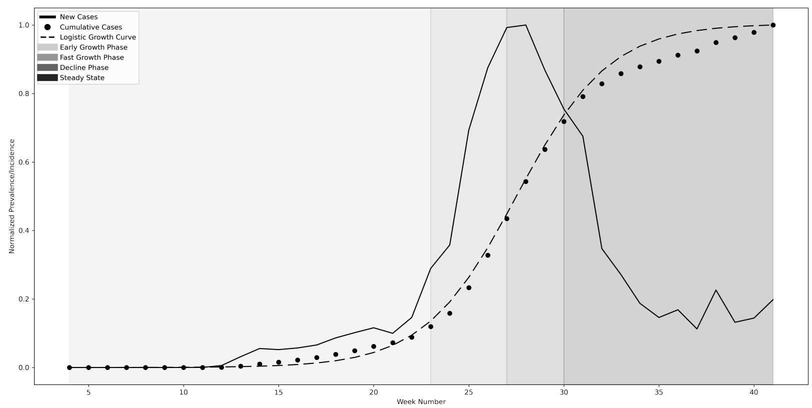

Figure 2. Logistic Growth Model applied to the state of Arizona

The incidence growth can be envisioned as transitions between various growth phases. Once the incidence curve goes into fast growth mode, the public health officials intervene to flatten the curve using warnings/outreach for people to stay home or promote face mask converting. The study takes the current and future incidence phases. The current phase is part of the antecedent, and the future phase is the consequent/target variable with a lag of 4 weeks. Based on a preliminary analysis, we found that the mobility, reopening mandates, and other factors are correlated with the number of cases with a lag of 4 weeks.

We collected 25 weeks of data, June 1, 2020 to November 15, 2020, across all 50 states. Since future incidence phase is lagged by 4 weeks, we end up with 21 weeks of transactional data. The dataset thus has 1050 transactions, with each transaction corresponding to 21 weeks, for each of the 50 states. An example rule would be, $MaskUsage : Statewide \& CurrentPhase : EarlyGrowth \Rightarrow FuturePhase : EarlyGrowth$. This rule implies that when a statewide mask mandate is active and the state is in the early growth phase, the state would continue to remain in the early growth phase. Mask usage, current phase, and future phase are the attributes. Statewide and early growth are the corresponding values for mask mandate and current incidence phase respectively. The antecedents in the dataset are mask mandates, state re-openings, mobility levels, case levels, population density, and current incidence rate. The consequent or the target variable is the future incidence rate. In this analysis, we set the minimum support threshold to 0.01, this means that the combination of factors in the antecedent and the consequent should appear in at least ten transactions (ten weeks of data) to be considered important. This threshold could mean that the antecedent can appear across 10 weeks in a single state or 1 week across 10 states or any combination in between. The minimum confidence is 0.7 and the minimum lift is 1.

Results

429 out of 55,125 relationships generated from the original transactions met the minimum threshold levels described in the Data and Methods section (support of 0.01, confidence of 0.7, and a lift value greater than 1). Each of these rules appeared in at least 10 transactions, i.e.,10 weeks of observations across the United States. With a confidence score of 0.7, each of the consequent (RHS) appears in at least 70% of the transactions with the antecedent (or the LHS). Finally, a high lift score (greater than 1) tells us that the factors in the antecedent were sufficiently positively correlated for deriving conclusions from the data.

Table 2 shows the top-5 association rules for various combinations of current and future incidence phases. These rules presented show various factors that contributed to the infection growth pattern, which is represented as one of four phases (i.e., early growth, fast growth, decline, and steady state). Of the 8 possible combinations between the current and the future incidence phases, we observe strong association rules that satisfy the minimum thresholds described above for 5 combinations: continued early growth, early growth to fast growth, continued fast growth, continued decline, and steady state to early growth. In Table 2, the first five rules highlight the circumstances where the incidence of cases stays constant, continuing in the same phase. The next five rules highlight scenarios where the incidence rate increases in the early growth phase and transitions into the fast growth phase. We also present the support, confidence, and lift values for each of these rules. These represent the coverage, strength, and predictive power of the rule respectively along with the chi-squared value of that rule. Given the maximum number of antecedents is 6, the critical-value of χ^2 is 11.0705 for a significance of $p<0.05$[35]. A chi-squared value greater than 11.07 implies that the rule is significant. All the association rules presented in Table 2 are significant.

We observed five combinations of current and future phases in the extracted association rules. The following are a summary of interesting observations:

1. *Continued Early Growth*: These rules represent the scenarios in which the number of cases continue to grow at a constant rate. The most important rule (i.e., 11% support and 97% confidence) shows that a state can remain in early growth rate even when there is a mask mandate. Another rule with lower support (5% support and 76% confidence) represents a scenario where states remain in the early growth rate without a mask mandate and high mobility. In addition, the rules in the continued early growth phase also demonstrate that states with a mask mandate, along with high mobility, medium-case levels, and phase-3 social distancing, will also continue in the early growth phase.

2. *Early Growth to Fast Growth*: Here, the number of cases increase rapidly leading to an explosion in the number of new cases. The top-5 rules that contributed to the fast growth phase from the early growth phase have no mask mandates as the underlying common factor. Moreover, these rules have strong support and high confidence when no-mask is combined with low mobility, strict social distancing guidelines (i.e., phase 0), and a low number of cases.

3. *Continued Fast Growth*: When a state is in a fast-growth phase, we did not observe a specific combination of factors that lead to a decrease in the number of cases.

4. *Continued decline*: When case counts were decreasing, the top-5 rules have either a county-level or a state-level mask mandate. This pattern is observed with multiple factors (high mobility, high case levels, and relaxed social distancing guidelines)

5. *Steady state to early growth*: When the states transitioned from a steady state to the early growth stage (indicating a resurgence in COVID 19 cases), we observed all the top-5 rules had a no-mask mandate. Other antecedents for these rules include a combination of lower number of cases, strict social distancing guidelines, and very high mobility.

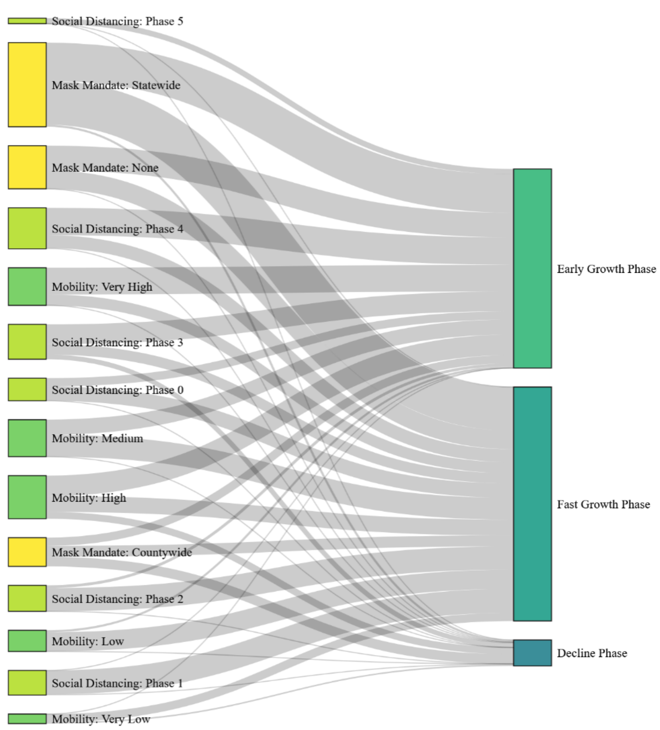

Figure 3. Association of various variables to the antecedent (future incidence curve of the pandemic)

We use a Sankey diagram to illustrate the combination of factors that contribute to different infection growth phases. The contributing factors are shown on the left, and the resulting phase from the combination of contributing factors is shown on the right. The width of the edge between the antecedent and the consequent represents the rules frequency for the given antecedent and consequent set. The flow lines show the relative strength of different factors (mask mandates, local mobility, population density, and social distancing orders) that contribute to the future incidence phase. The higher the number of rules for a particular variable, the larger the impact of that variable in affecting the outcome in the incidence. For example, in the case

Table 2. Top 5 association rules for different combinations of current and future phases, their support, confidence, lift, and chi-squred measures

Association Rule	Support	Confidence	Lift	Chisquared
Continued Early Growth				
Mask Usage: Statewide Current Phase: Early Growth ⇒ Future Phase: Early Growth	0.11	0.97	2.76	225.44
Mask Usage: Statewide Social Distancing: Phase 3 Current Phase: Early Growth ⇒ Future Phase: Early Growth	0.05	0.93	2.65	87.82
Mask Usage: None Mobility Levels: Very High Current Phase: Early Growth ⇒ Future Phase: Early Growth	0.05	0.76	2.17	54.56
Mask Usage: Statewide Case Level: Medium Current Phase: Early Growth ⇒ Future Phase: Early Growth	0.04	1.00	2.85	80.94
Mask Usage: Statewide Mobility Level: High Current Phase: Early Growth ⇒ Future Phase: Early Growth	0.04	0.96	2.73	74.11
Early Growth to Fast Growth				
Mask Usage: None Case Level: Very Low Current Phase: Early Growth ⇒ Future Phase: Fast Growth	0.02	0.71	1.85	13.69
Mask Usage: None Case Level: Very Low Social Distancing: Phase 0 Current Phase: Early Growth ⇒ Future Phase: Fast Growth	0.02	0.75	1.96	16.43
Mask Usage: None Case Level: Very Low Social Distancing: Phase 0 Population Density: Low Current Phase: Early Growth ⇒ Future Phase: Fast Growth	0.02	0.73	1.90	14.95
Mask Usage: None Mobility Level: Very Low Current Phase: Early Growth ⇒ Future Phase: Fast Growth	0.01	1.00	2.62	17.18
Mask Usage: None Mobility Level: Very Low Social Distancing: Phase 0 Current Phase: Early Growth ⇒ Future Phase: Fast Growth	0.01	1.00	2.62	17.18
Continued Fast Growth				
Mask Usage: Statewide Current Phase: Fast Growth ⇒ Future Phase: Fast Growth	0.16	0.78	2.04	181.43
Mask Usage: Statewide Population Density: Low Current Phase: Fast Growth ⇒ Future Phase: Fast Growth	0.12	0.79	2.08	134.33
Mask Usage: None Current Phase: Fast Growth ⇒ Future Phase: Fast Growth	0.10	0.72	1.89	82.55
Mask Usage: None Population Density: Low Current Phase: Fast Growth ⇒ Future Phase: Fast Growth	0.10	0.72	1.89	82.55
Mask Usage: Statewide Social Distancing: Phase 1 Current Phase: Fast Growth ⇒ Future Phase: Fast Growth	0.06	0.76	1.99	54.5
Continued Decline				
Mask Usage: Countywide Current State: Decline ⇒ Future State: Decline	0.03	0.85	4.12	97.2
Mask Usage: Countywide Population Density: Low Current State: Decline ⇒ Future State: Decline	0.03	0.94	4.54	113.27
Mask Usage: Countywide Population Density: Low Social Distancing: Phase 3 Current State: Decline ⇒ Future State: Decline	0.02	1.00	4.85	82.5
Mask Usage: Countywide Social Distancing: Phase 3 Current State: Decline ⇒ Future State: Decline	0.02	0.83	4.04	61.95
Mask Usage: Countywide Case Level: High Current State: Decline ⇒ Future State: Decline	0.02	0.82	3.97	60.28
Steady State to Early Growth				
Mask Usage: None Current Phase: Steady State ⇒ Future Phase: Early Growth	0.03	0.90	2.65	50.69
Mask Usage: None Population Density: Low Current Phase: Steady State ⇒ Future Phase: Early Growth	0.03	0.90	2.69	51.99
Mask Usage: None Mobility Level: Very High Social Distancing: Phase 0 Current Phase: Steady State ⇒ Future Phase: Early Growth	0.02	1.00	2.85	39.64
Mask Usage: None Social Distancing: Phase 0 Current Phase: Steady State ⇒ Future Phase: Early Growth	0.02	0.94	2.70	35.23
Mask Usage: None Mobility Level: Very High Current Phase: Steady State ⇒ Future Phase: Early Growth	0.02	1.00	2.85	39.64

of statewide face mask mandate, the highest number of rules (77 rules) are associated with the early growth phase, followed by the fast growth phase (66 rules), and the declining phase has the least number of rules (12 rules) in the dataset. The following are some interesting observations from Figure 2.

- Rules with no mask mandate are only associated with an early growth (54.34%) or a fast growth phases (45.65%). There were no rules with a no mask mandate where the future incidence phase is a decline phase or a steady state phase.
- In comparison, the rules with mask mandates (statewide and countywide) were associated with all three future incidence phases: early growth, fast growth, and decline phases with 52.12%, 35.1%, and 12.76% rules in each phase respectively.
- Reopening guidelines issued by the states were strongly associated with specific phases of the pandemic. Strict guidelines instituted during Phase 0 were always associated with rules in the early growth and the fast growth phases, as most states imposed strict lock-downs as the number of cases started to increase. On the other hand, the incidence of cases increased when these restrictions were relaxed. Phase 3 and 4 reopening guidelines led to a resurgence in the incidence(early growth and fast growth) in 87.74% of the rules, and a decrease in incidence was observed in 12.24% of the rules.
- Mobility has a considerable impact in determining the future phase of the pandemic. Lower mobility is associated with the early growth phase, 3.2% of the total rules associated with low or very low mobility compared with 80.6% of rules leading to a fast growth phase, and 16.12% of rules where the future phase is a decline phase. On the other hand, the rules with medium or higher mobility were associated mainly with future phases leading to early growth, fast growth, and decline phases 65.7%, 30.09%, and 3.3% respectively. These distributions imply that lower mobility is associated with a decline in the number of cases, while higher mobility is associated with resurgence and an increase in the number of cases.

Discussion

COVID-19 policies with respect to mobility restriction, shutdowns, mask mandates, etc., are currently the nation's highest priorities with respect to saving lives and protecting the economy. Identifying and profiling the combination of policies that worked and did not work is important. This provides the necessary data for a rational decision support framework on how best to manage policies at the state level, given their diverse attributes. While the existing studies provide individual correlations, associations, forecasting, etc., they do not provide insights into effective combinations. The goal of our proposed method is to improve this understanding to aid policymakers in making the right decisions to help minimize spread while balancing convenience and economic growth priorities.

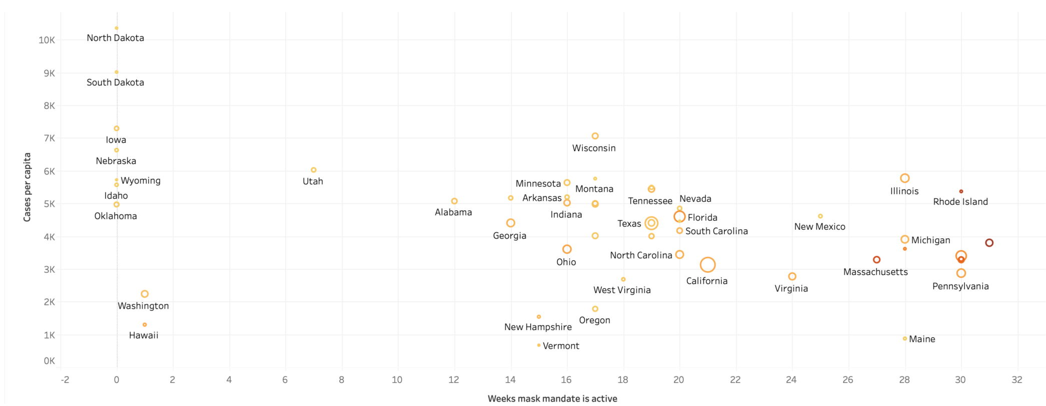

Figure 4. Relationship between the number of cases per capita and the number of weeks a mask mandate is active in a state

Relationship between case-growth and mask mandates

Based on the association rules in Table 2, no mask mandates were always associated with an increase in the number of cases and mask mandates were associated with a decrease in the number of cases. While it is not clear which specific measures led to a decrease in the number of cases, the mask mandates were always associated with a continued decline in the number of new cases. Most of the states issued a mask mandate when the number of cases was increasing rapidly, alongside stay-at-home orders. This observation is in line with earlier research showing that strong social distancing measures reduced the number of cases [36]. However, the effect of mask mandates separate from social distancing measures is not apparent in the fast growth phase. This was because the two measures were typically instituted together when the cases were increasing. For this reason, we cannot assess differential contributions of these measures. We observed that the mask mandates were effective in the early growth and decline phases of the pandemic. We also observed that the states that did not institute a mask mandate continued to see an increase in the number of cases for a longer duration than the states that did. Figure 4 shows the relationship between the number of cases per capita, and the length of time the mask mandates were active in the different states. The color of the map shows the population density of a state and the size shows the number of cases in that state. We observe that the longer the duration for which the mask mandates were active, the lower were the number of cases per capita. We also observed that states with high population densities that instituted a mask mandate had a lower number of cases per capita.

Relationship between mobility and case-growth

Our results from Table 2 and Figure 3 indicate that mobility also impacts the incidence rate of the pandemic. The association rules indicate that increased mobility and a lack of mask mandates were associated with a resurgence of cases. A majority of the states in the United States were able to successfully control the spread of the pandemic in spring and summer with strict social distancing guidelines and the resultant reduction in mobility. However, all the states had an increase in the number of cases in October and November, despite having issued mask mandates at state and county levels. This was likely related to increased mobility during this time period. In states that did not institute mask mandates, there was an increase in the number of cases irrespective of the mobility levels or the social distancing guidelines issued by the state and local authorities. By this, we surmise that social distancing and masking regulations were by themselves inadequate to reduce the number of new cases.

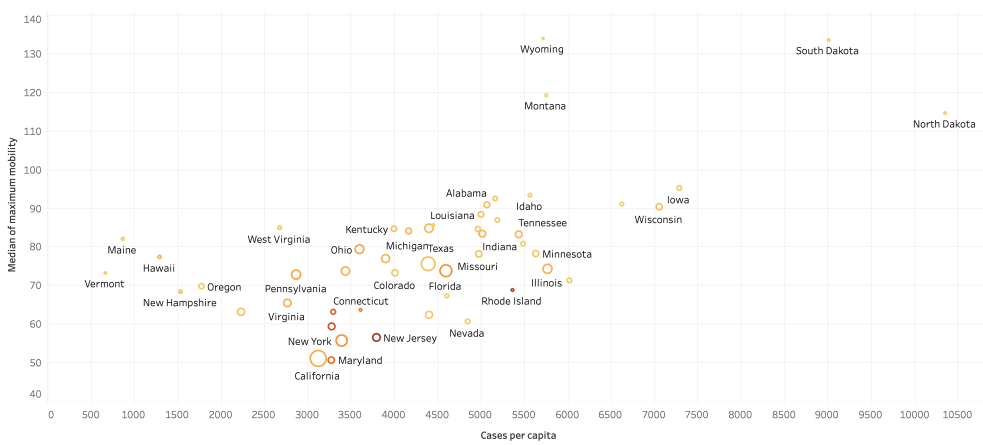

Figure 5. Impact of number of weeks mask mandate was active and the mobility on the number of cases per capita

Figure 5 shows the relationship between the number of cases per capita and the median of maximum mobility for that state at a weekly level of granularity. The size of each marker shows the total number of cases and the color indicates the number of weeks that state has a mask mandate. The states with mobility lower than 80 percent of the baseline had a lower number of cases per capita compared to states that had higher mobility. The states with the highest mobility, i.e., South Dakota, North Dakota, Wyoming, and Montana were also the states with a considerably higher number of cases. These observations indicate that while mask mandates are essential, reducing the mobility of individuals and strict regulations on the businesses open also had a significant association with a reduction in the number of cases.

The states that did not institute mask mandates did not also impose strict social distancing guidelines or relaxed the guidelines earlier than most of the other states. These include states like South Dakota, Mississippi, North Dakota, and Utah. Both North Dakota and Utah imposed strict statewide mask mandates in mid-November when the number of cases increased exponentially. Our results in Table 2 and Figure 3 show the effect that various mask mandates, socials distancing guidelines, and mobility had on the change in the growth rate of the pandemic.

Limitations and Future Work

We emphasize the limited scope of our analysis, as it is important to interpret these results with a clear understanding of the limitations with respect to both the data quality and the methodology.

Our data includes the start and end dates of various interventions by state and local authorities; but this does not help us measure the actual compliance to these measures. With respect to the mask mandates issued at a county level, in a majority of the states, the population under the coverage of the mandates or recommendations is not known. We also did not consider several other conditions that affect growth in cases. For example, the analysis does not consider events such as holidays, weather conditions, congregation events, etc. Our assumptions about the incidence growth phase were based on the best fit from the logistic growth model.

In ARM, the choice of parameters (i.e., support and confidence thresholds) affect the rules generated[36]. If the thresholds are set too high, then we obtain very few rules. If the thresholds are set too low, we obtain too many rules. To make the analysis less susceptible to thresholds, we use the top-5 rules to study the impact of various factors to account for changes in phases of the pandemic. The discretization of variables also affects the type of rules generated. For instance, using just 3 classes (low, medium, and high) rather than 5 classes (very low, low, medium, high, and very high) produces a very different set of rules. We use five-class categorization using symmetric quantiles to cope with the discretization issue; and this approach would provide an optimal analysis for association mining. In the future, a supervised discretization technique based on the strength of association rules can be used to further improve the quality of the rules generated. Future work can explore sensitivity analysis towards this goal. This approach provides a new direction to develop AI-based techniques that can provide policy recommendations for policymakers on various actions could potentially decrease the number of new cases.

Conclusion

The paper introduces a novel approach to analyze the effects of different non-pharmaceutical interventions to contain and manage the infection growth rate. The proposed approach uses the Association Rule Mining technique and discretization of infection growth phases using multi-peak logistic growth model. We noted several interesting observations. For instance, there is strong similarity between states that had strict mask mandates and reduced infection growth rate. Also, no difference was observed in terms of infection growth rate between state-wide versus county wide mask mandates. Various other factors such as population density, and mobility levels impacted the increase in the number of cases, highlighting the importance of local factors on the number of COVID cases. These findings are important as the United States is trying to reach herd immunity through vaccination, while balancing against a relaxation of the social distancing guidelines.

References

1. Fernandes, N. Economic effects of coronavirus outbreak (covid-19) on the world economy. *Available at SSRN 3557504* (2020).
2. Gapen, M., Millar, J., Blerina, U. & Sriram, P. Assessing the effectiveness of alternative measures to slow the spread of covid-19 in the united states. *Covid Econ.* **40**, 46–75 (2020).
3. Li, J. *et al.* Do stay at home orders and cloth face coverings control covid-19 in new york city? results from a sier model based on real-world data. In *Open Forum Infectious Diseases*, vol. 8, ofaa442, DOI: 10.1093/ofid/ofaa442 (Oxford University Press US, 2021).
4. Tatapudi, H., Das, R. & Das, T. K. Impact assessment of full and partial stay-at-home orders, face mask usage, and contact tracing: An agent-based simulation study of covid-19 for an urban region. *Glob. Epidemiol.* **2**, 100036, DOI: 10.1016/j.gloepi.2020.100036 (2020).
5. Silva, P. C. *et al.* Covid-abs: An agent-based model of covid-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos, Solitons & Fractals* **139**, 110088, DOI: 10.1016/j.chaos.2020.110088 (2020).
6. Ghaffarzadegan, N. Simulation-based what-if analysis for controlling the spread of covid-19 in universities. *Plos one* **16**, e0246323, DOI: 10.1371/journal.pone.0246323 (2021).
7. Badr, H. S. *et al.* Association between mobility patterns and covid-19 transmission in the usa: a mathematical modelling study. *The Lancet Infect. Dis.* **20**, 1247–1254, DOI: "https://doi.org/10.1016/S1473-3099(20)30553-3" (2020).
8. Sarmadi, M., Marufi, N. & Moghaddam, V. K. Association of covid-19 global distribution and environmental and demographic factors: An updated three-month study. *Environ. Res.* **188**, 109748, DOI: 10.1016/j.envres.2020.109748 (2020).
9. Agrawal, R., Imieliński, T. & Swami, A. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, 207–216, DOI: doi.org/10.1145/170036.170072 (1993).
10. Brossette, S. E. *et al.* Association rules and data mining in hospital infection control and public health surveillance. *J. Am. medical informatics association* **5**, 373–381, DOI: 10.1136/jamia.1998.0050373 (1998).
11. Paetz, J. & Brause, R. A frequent patterns tree approach for rule generation with categorical septic shock patient data. In *International Symposium on Medical Data Analysis*, 207–213, DOI: 10.1007/3-540-45497-7_31 (Springer, 2001).

12. Chen, J., He, H., Williams, G. & Jin, H. Temporal sequence associations for rare events. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 235–239, DOI: 10.1007/978-3-540-24775-3_30 (Springer, 2004).

13. Ordonez, C., Ezquerra, N. & Santana, C. A. Constraining and summarizing association rules in medical data. *Knowl. information systems* **9**, 1–2, DOI: 10.1007/s10115-005-0226-5 (2006).

14. Huang, H., Tornero-Velez, R. & Barzyk, T. M. Associations between socio-demographic characteristics and chemical concentrations contributing to cumulative exposures in the united states. *J. exposure science & environmental epidemiology* **27**, 544–550, DOI: doi.org/10.1038/jes.2017.15 (2017).

15. Kadi, N. & Khelfaoui, M. Population density, a factor in the spread of covid-19 in algeria: statistic study. *Bull. Natl. Res. Centre* **44**, 1–7 (2020).

16. Bhadra, A., Mukherjee, A. & Sarkar, K. Impact of population density on covid-19 infected and mortality rate in india. *Model. Earth Syst. Environ.* **7**, 623–629 (2021).

17. Feng, S. *et al.* Rational use of face masks in the covid-19 pandemic. *The Lancet Respir. Medicine* **8**, 434–436 (2020).

18. Sen, S., Karaca-Mandic, P. & Georgiou, A. Association of stay-at-home orders with covid-19 hospitalizations in 4 states. *Jama* **323**, 2522–2524 (2020).

19. Shimada, K., Hirasawa, K. & Hu, J. Class association rule mining with chi-squared test using genetic network programming. In *2006 IEEE International Conference on Systems, Man and Cybernetics*, vol. 6, 5338–5344 (IEEE, 2006).

20. Alvarez, S. A. Chi-squared computation for association rules: preliminary results. *Boston, MA: Boston Coll.* **13** (2003).

21. Xu, Y., Zhou, S.-X. & Gong, J.-H. Mining association rules with new measure criteria. In *2005 International Conference on Machine Learning and Cybernetics*, vol. 4, 2257–2260 Vol. 4, DOI: 10.1109/ICMLC.2005.1527320 (2005).

22. Schuchat, A., Covid, C. & Team, R. Public health response to the initiation and spread of pandemic covid-19 in the united states, february 24–april 21, 2020. *Morb. Mortal. Wkly. Rep.* **69**, 551, DOI: 10.15585/mmwr.mm6918e2externalicon (2020).

23. Markowitz, A. State-by-state guide to face mask requirements. *AARP. Retrieved online on Febr.* **10** (2021). Available at https://gtxcorp.com/aarp-com-state-by-state-guide-to-face-mask-requirements.

24. Masks4All. What u.s. states require masks in public? https://masks4all.co/what-states-require-masks/ (2021). (Date of Access: 2021-01-20).

25. Katragadda, S. Github: Association mining - data collection and preprocessing (2021). URL https://github.com/raviteja-bhupatiraju/AssociationMining_COVID19.

26. Ballotpedia. State government responses to the coronavirus. https://ballotpedia.org/State_government_responses_to_the_coronavirus_(COVID-19)_pandemic,_2020 (2020). (Date of Access: 2020-12-25).

27. The Food Industry Association. Covid-19 - state reopening plans. https://www.fmi.org/blog/view/state-affairs-issue-papers/2020/12/08/covid-19---state-reopening-plans (2020). (Date of Access: 2020-12-25).

28. Kang, Y. *et al.* Multiscale dynamic human mobility flow dataset in the us during the covid-19 epidemic. *Sci. data* **7**, 1–13, DOI: 10.1038/s41597-020-00734-5 (2020).

29. Pan, Y. *et al.* Quantifying human mobility behaviour changes during the covid-19 outbreak in the united states. *Sci. Reports* **10**, 1–9, DOI: doi.org/10.1038/s41598-020-77751-2 (2020).

30. Warren, M. S. & Skillman, S. W. Mobility changes in response to covid-19. *arXiv preprint arXiv:2003.14228* (2020). Preprint at https://arxiv.org/abs/2003.14228.

31. World Population Review. United states by density 2021. https://worldpopulationreview.com/state-rankings/state-densities (2020). (Date of Access: 2020-12-25).

32. Johns Hopkins University. Coronavirus resource center. https://coronavirus.jhu.edu/ (2020). (Date of Access: 2021-01-15).

33. Batista, M. Estimation of the final size of the second phase of coronavirus epidemic by the logistic model (2020). Preprint at https://doi.org/10.1101/2020.03.11.20024901.

34. Wu, K., Darcet, D., Wang, Q. & Sornette, D. Generalized logistic growth modeling of the covid-19 outbreak: comparing the dynamics in the 29 provinces in china and in the rest of the world. *Nonlinear dynamics* **101**, 1561–1581, DOI: https://doi.org/10.1007/s11071-020-05862-6 (2020).

35. Kokoska, S. & Nevison, C. Critical values for the chi-square distribution. In *Statistical Tables and Formulae*, 58–59 (Springer, 1989).

36. García, M. N. M., Román, I. R., Peñalvo, F. J. G. & Bonilla, M. T. An association rule mining method for estimating the impact of project management policies on software quality, development time and effort. *Expert. Syst. with Appl.* **34**, 522–529 (2008).

Acknowledgements

This research was partially funded by NSF grants CNS-1650551, CNS-2027688, and CNS-1429526.

Author contributions statement

Problem description: R.G and R.K; Conceptualization: R.G.; Methodology: S.K.; Software: S.K. and A.K.; Validation: S.K. and R.G.; Formal analysis: S.K., V.R., and R.G.; Investigation: R.B. and Z.A.; Resources: R.G.; Data curation: R.B., S.K. and R.G.; Writing—original draft preparation: S.K., R.G., and R.B.; Writing—review and editing: Z.A. and V.R.; Visualization: R.B.; Supervision: R.G.; Project administration: R.G.; Funding acquisition: R.G., V.R., H.C., and R.K. All authors reviewed the manuscript.

Additional information

The authors report no competing interests. The analysis code for this paper is available on GitHub at https://github.com/raviteja-bhupatiraju/AssociationMining_COVID19